(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 642 116 A1**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **94113017.1**

(22) Date of filing: **20.08.94**

(51) Int. Cl.[6]: **G10L 3/00**, G10L 5/00, G10L 7/02, G10L 9/06

(30) Priority: **06.09.93 IT MI931905**

(43) Date of publication of application: **08.03.95 Bulletin 95/10**

(84) Designated Contracting States: **BE DE ES FR GB NL SE**

(71) Applicant: **ALCATEL ITALIA S.p.A.**
Via L. Bodio, 33/39
I-20158 Milano (IT)

(72) Inventor: **Riccio, Antonello**
Via Premuda 34
I-84132 Salerno (IT)
Inventor: **Di Ronza, Benedetto**
Via Guido de Ruggiero 50
I-70125 Bari (IT)

(74) Representative: **Pohl, Herbert, Dipl.-Ing et al**
Alcatel SEL AG
Patent- und Lizenzwesen
Postfach 30 09 29
D-70449 Stuttgart (DE)

(54) Method of generating components of a speech database using the speech synthesis technique and machine for automatic speech recognition.

(57) The invention relates to a method of generating components of a speech database using the speech synthesis technique and to a machine for automatic speech recognition.

Asking a speaker to repeat speech elements on the basis of a preventive series of utterances corresponding to automatically synthesized utterances of said speech elements, the quality of the database thus obtained is better and uniform.

EP 0 642 116 A1

The present invention relates to a method of generating components of a speech database using the speech synthesis technique and to a machine for automatic speech recognition.

Basically, machines for speech recognition can be divided into two categories: the first ones are based upon conventional processors and realize the recognition by comparing the word to be recognized with words of a pre-established vocabulary; the second ones are based on special architectures such as neural networks and the pre-established vocabulary depends on a set of parameter values characterizing such architectures; hence the first machines require the generation of a speech database corresponding to the pre-established vocabulary for their operation, while the second ones require the generation of a suitable set of parameter values corresponding to the pre-established vocabulary which could be considered as a distributed speech database.

As known, the generation of such databases, either concentrated or distributed, occurs through long and repeated recording operations; in general, a team of speakers is selected and each speaker utters a number of times the speech elements corresponding to a predetermined vocabulary (in general words or, less frequently, the syllables); the acoustic signals corresponding to such utterance are acquired and often tape-recorded; subsequently a processing step may follow consisting, e.g., in a background noise filtering, in a sampling and in a digitizing; lastly, the database real generation is carried out, which may simply consist in storing on semiconductor storages according to a pre-established format or, in addition to and before storage, in the generation of suitable parameters, for instance LPC (Linear Predictive Code), starting from the acquired and processed acoustic signals ; in case of neural networks, the generation of the distributed database occurs by directly providing the network (that subsequently will carry out the recognition) with the acquired and processed acoustic signals and by leaving the network itself changing the values of its parameters during a step called "training".

The word "training", when referred to machines for speech recognition belonging to the first category, indicates an operative step during which the concentrated database is enhanced with new utterances of speech elements belonging or less to the predetermined vocabulary; not all such machines feature a "training" step.

The generation of such databases must be realized with great care since the recognition rate strongly depends on the used data base.

Substantially there are two methods by which the speakers can be allowed to utter the speech elements belonging to the predetermined vocabulary.

The first one consists in providing each speaker with a written list of the speech elements to be uttered: this method has the disadvantage, very heavy if speakers are unprofessional, of leading to an unnatural and erratic pronunciation due to the fact that the speaker starts the utterance with some voice characteristics, such as high energy, considerably imperative prosody, high speed of words alternated each other by a long silence, clear and well scanned articulation of syllables, and terminates the utterance with lower energy, more apathetic (i.e. meaningless) prosody, low speed of words alternated each other by a short silence, fluent articulation of syllables. Such a method, moreover, is not always applicable as in case of acquisition of speech databases from telephone line, where the speaker is chosen at random among the subscribers, or in a car where the driver cannot drive and read at the same time.

The second method can be applied, e.g., in the just mentioned cases and consists in asking the speaker to repeat the speech elements uttered by another people called "operator": it has been discovered that, with such method, the utterance of the speaker is disadvantageously altered because the speaker tends to "copy" the pronunciation of the operator, more precisely: speed and energy of words and articulation of their components, accentuation of vowels, prosody of words, emphasis, cadence or rhythm of syllables and eventual personal characteristics of the operator (e.g. dialect, emotional, physical characteristics). If the operator or his personal characteristics change, as is the frequent case of prolonged recording operations, a totally uneven database will be obtained and this is a further disadvantage.

The object of the present invention is to overcome the drawbacks of the known technique.

This object is achieved through the method of generating a component of a speech database as set forth in claim 1, through the methods of generating and enhancing a speech database as set forth in claims 8 and 9 respectively, and through the machine for automatic speech recognition as set forth in claim 10.

Further advantageous aspects of the present invention are set forth in the subclaims.

By asking the speaker to repeat speech elements on the basis of a preventive series of voiced emissions corresponding to automatically synthesized utterances of such speech elements, the quality of the database thus obtained is better and uniform.

If such synthetic utterances are particularly unnatural, the copy-effect is limited.

Moreover, by making the sound emission be cadenced in a pre-established manner by a ma-

chine, the list effect is limited.

Lastly, by making such synthetic utterances be automatically synthesized still in accordance with the same method and the same synthesis parameters, the uniformity of the database is considerably improved.

The present invention will result better from the following description.

In accordance with the present invention, the method of generating a component of a speech database corresponding to the utterance of a speech element, comprises the steps of :

a) emitting an automatically synthesized utterance of the speech element,

b) waiting for a speech acoustic signal corresponding to an utterance of the speech element, and

c) acquiring such speech acoustic signal.

Such synthesized utterance can be synthesized starting from a text or from a natural pronunciation so modified to be unnatural, in particular prosodically unnatural.

Such acquired speech acoustic signal conceptually corresponds to the desired component of the speech database.

It has been said "conceptually" because the component assumes very different forms depending on the circumstances: for instance, in the case of a neural network speech database, it will be formed by some values of network parameters or by some variations of the same not connected in a simple way to the acquired acoustic signal and calculated automatically by the network itself according to a predetermined algorithm.

Then, such acquired speech acoustic signal is processed: sometimes by carrying out a simple sampling and digitizing sometimes through complicated digital coding algorithms, still sometimes through analog operations.

The generation of a speech database component can be carried out either prior to the speech recognition step from equipments arranged on purpose or during the same step from the same machine for speech recognition.

In the first case, the acquisition step is almost always overseen by an operator who takes care at least of the recording operations and who has the possibility of recognizing the occurrence of anomalous conditions and find a remedy for them.

In particular, in the second case, it is to advantage that the method further comprises the step of :

d) verifying the acquired speech acoustic signal corresponds to an utterance of the voice element, e.g., through comparison with the synthesized utterance.

Should this verification be unsuccessful, such anomalous condition can be signalled.

Moreover, it may be advisable to provide that, if step b) is longer than a predetermined period of time, step c) does not take place; hence the generation of the element has failed and such anomalous condition can be signalled.

Such signallings are useful to the speaker who is uttering the speech element of a vocabulary in order that he can find a remedy for them.

The fact that the utterance is synthetic can be advantageously utilized in two different ways: if the pronunciation is greatly unnatural, the speaker will be neither able nor inclined to "copy" such pronunciation (consequently the synthesizer will be very easy to realize and cheaper); if on the other hand the synthetic pronunciation is fairly natural, the speaker will be inclined to copy it, of course; therefore, it can be thought to individuate, through laboratory tests, those values of the synthesis parameters which allow the achievement of an "ideal" pronunciation from the speaker that is to say which provides the best results during recognition step; in any case the possibility of varying the synthesis parameters will allow the indirect control of the speaker's pronunciation advantageously.

Step a) can be preceded by a step of taking the synthetic pronunciation out of storage means or out of an automatic synthesis step of such pronunciation starting from the corresponding speech element.

Methods of generating and increasing speech databases can be derived on the basis of the method of generating only one component of a speech database, just described.

In accordance with the present invention, the method of generating a speech database corresponding to a predetermined vocabulary comprising a plurality of speech elements provides that the steps of the method just described are repeated at least once for each speech element of the vocabulary.

According to the present invention, the method of enhancing a speech database corresponding to utterances of speech elements belonging to a predetermined vocabulary, through new utterances of speech elements belonging or less to such vocabulary, provides that the steps of the method just described are repeated for each new utterance.

It has been explained how the machines for automatic speech recognition provide a "training" operative step during which the speech database corresponding to a recognition vocabulary of speech elements is generated or enhanced.

The machine for automatic speech recognition, in accord to the present invention, during the training step is designed to prepare itself in order to realize the steps of the method described above.

In a particularly simple embodiment, such machine comprises storage means capable of contain-

ing automatically synthesized utterances of speech elements of the recognition vocabulary.

In this circumstance, it will be enough to read out from such storage means the various utterances, emit them one at a time and wait for the reciter to repeat them.

Such storage means may coincide with those for containing the speech database.

Starting from an extremely simple initial speech database having only one component for each speech element of the vocabulary, such database can be enlarged by using the utterances of the initial database as synthetic utterances: if the first ones come from automatic synthesis operations they can be emitted without further processing, otherwise they can be modified in such a way as to result unnatural, as already said.

In a second embodiment, such machine comprises an automatic speech synthesizer capable of synthesizing and emitting utterances of speech elements also not comprised in the recognition vocabulary.

The recognition vocabulary is chosen during production; it may be contemplated that the purchaser personalizes the vocabulary by adding personal speech elements.

Such personal speech elements may, e.g., be introduced into such machine in a textual form and synthesized and emitted by the synthesizer during the training step.

Claims

1. Method of generating a component of a speech database corresponding to the utterance of a speech element, comprising the steps of :
   a) emitting an automatically synthesized utterance of said speech element,
   b) waiting for a speech acoustic signal corresponding to an utterance of said speech element, and
   c) acquiring said speech acoustic signal; whereby such acquired speech acoustic signal conceptually corresponds to said component.

2. Method according to claim 1, characterized in that said acquired speech acoustic signal is processed.

3. Method according to claim 1, characterized in that it further comprises the step of
   d) verifying that said acquired speech acoustic signal corresponds to an utterance of said speech elements through comparison with said synthesized utterance.

4. Method according to claim 1, characterized in that, if said step b) is longer than a predetermined period of time, said step c) does not take place and such anomalous condition is signalled.

5. Method according to claim 1, characterized in that said synthesized utterance is of unnatural type.

6. Method according to claim 1, characterized in that said step a) is preceded by a step for extracting said utterance from storage means.

7. Method according to claim 1, characterized in that said step a) is preceded by a step of automatic synthesizing said utterance starting from the corresponding speech element.

8. Method of generating a speech database corresponding to a predetermined vocabulary comprising a plurality of speech elements, characterized in that said steps of the method of claim 1 are repeated at least once for each speech element of said vocabulary.

9. Method of enhancing a speech database corresponding to utterances of speech elements belonging to a predetermined vocabulary, through new utterances, of speech elements belonging or less to said vocabulary, characterized in that the steps of the method of claim 1 are repeated for each new utterance.

10. Machine for the automatic recognition of speech in relation to a predetermined recognition vocabulary of speech elements, characterized in that, during the step of training, it is capable of arranging itself in such a way as to realize the steps of the method of claim 1.

11. Machine according to claim 10, characterized in that it comprises storage means designed to contain automatically synthesized utterances of speech elements of said recognition vocabulary.

12. Machine according to claim 10, characterized in that it comprises an automatic speech synthesizer designed to synthesize and emit utterances of speech elements even not comprised in said recognition vocabulary.

European Patent Office

EUROPEAN SEARCH REPORT

Application Number EP 94113017.1

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl. 6)
X	GB - A - 2 165 969 (BRITISH TELECOMMUNICATIONS PLC) * Fig. 1,2; abstract; claim 1 *	1	G 10 L 3/00 G 10 L 5/00 G 10 L 7/02 G 10 L 9/06
A	EP - A - 0 518 638 (TEXAS INSTRUMENTS INC.) * Abstract; claim 1 *	1	
A	EP - A - 0 451 695 (TEXAS INSTRUMENTS INC.) * Fig. 1; abstract; claim 1 *	1	

TECHNICAL FIELDS SEARCHED (Int. Cl. 6)

G 10 L 3/00
G 10 L 5/00
H 03 M 3/00
G 06 F 3/00
G 06 F 15/00
G 10 L 7/00
G 10 L 9/00
H 04 Q 3/00

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
VIENNA	29-11-1994	BERGER

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P0401)